

What Is Claimed Is:

1. A method of processing traffic received from an InfiniBand node via a first queue pair, comprising:

5 selecting a traffic entry in an InfiniBand receive queue, wherein said traffic entry comprises one of:

a Send command comprising an encapsulated communication;

a Send command comprising an RDMA Read descriptor; and

an RDMA Read response comprising a response to an RDMA

10 Read request;

if said selected traffic entry comprises a Send command comprising an RDMA Read descriptor:

issuing a first RDMA Read request to retrieve one or more portions of a communication described by said RDMA Read descriptor;

15 in a linked list corresponding to the first queue pair, adding an entry corresponding to said first RDMA Read request, said entry identifying a range of sequence numbers associated with expected responses to said first RDMA Read request; and

in a retry queue, adding an entry corresponding to said first RDMA Read request; and

20 if said selected traffic entry comprises an RDMA Read response to said first RDMA Read request:

identifying a sequence number associated with said RDMA Read response;

25 comparing said sequence number to said range of sequence numbers;

storing said one or more portions of said described communication

to facilitate assembly of said described communication in said queue; and
if said sequence number matches a final sequence number in said
range, retiring in said retry queue said entry corresponding to said first
RDMA Read request.

5

2. The method of claim 1, further comprising:

forwarding a communication associated with said selected traffic entry, for
transmission on an external communication link, wherein said communication is
one of:

10

said encapsulated communication; and
said described communication, after said described communication
is assembled.

3. The method of claim 1, further comprising, if said selected traffic
15 entry comprises an RDMA Read response to said first RDMA Read request:
if said sequence number does not match said final sequence number,
updating said entry in said linked list to include said sequence number.

4. The method of claim 1, further comprising:
20 maintaining a single memory structure comprising multiple linked list,
including said linked list;
wherein each linked list stores entries associated with RDMA Read
requests for a different InfiniBand queue pair.

25 5. The method of claim 1, further comprising:
maintaining a single memory structure for queuing InfiniBand traffic
received via multiple virtual lanes and multiple queue pairs, said single memory

structure comprising said queue.

6. The method of claim 5, wherein said queue comprises a linked list of memory buffers within said single memory structure.

5

7. The method of claim 1, further comprising:
maintaining a head pointer configured to identify a head of said linked list;
and
maintaining a tail pointer configured to identify a tail of said linked list.

10

8. The method of claim 1, further comprising:
maintaining a head pointer configured to identify a head of said queue;
maintaining a tail pointer configured to identify a tail of said queue; and
maintaining a next traffic entry pointer configured to identify a next entry
15 in said queue to be processed after said forwarding.

9. The method of claim 8, wherein said tail pointer is configured to identify where in said queue a next traffic entry is to be queued.

20 10. The method of claim 1, further comprising, if said selected traffic entry comprises an RDMA Read descriptor:
appending space to a head of said queue;
wherein said described communication is assembled in said appended space.

25

11. The method of claim 1, further comprising, if said selected traffic entry comprises an RDMA Read response to said first RDMA Read request:

dropping an RDMA Read response received out of order; and
requesting a retry of said first RDMA Read request.

12. A computer readable medium storing instructions that, when
5 executed by a computer, cause the computer to perform a method of processing
traffic received from an InfiniBand node via a first queue pair, the method
comprising:

selecting a traffic entry in an InfiniBand receive queue, wherein said traffic
entry comprises one of:

10 a Send command comprising an encapsulated communication;
a Send command comprising an RDMA Read descriptor; and
an RDMA Read response comprising a response to an RDMA
Read request;

if said selected traffic entry comprises a Send command comprising an
15 RDMA Read descriptor:

issuing a first RDMA Read request to retrieve one or more portions
of a communication described by said RDMA Read descriptor;

in a linked list corresponding to the first queue pair, adding an
entry corresponding to said first RDMA Read request, said entry
20 identifying a range of sequence numbers associated with expected
responses to said first RDMA Read request; and

in a retry queue, adding an entry corresponding to said first RDMA
Read request; and

if said selected traffic entry comprises an RDMA Read response to said
25 first RDMA Read request:

identifying a sequence number associated with said RDMA Read
response;

comparing said sequence number to said range of sequence numbers;

storing said one or more portions of said described communication to facilitate assembly of said described communication in said queue; and

5 if said sequence number matches a final sequence number in said range, retiring in said retry queue said entry corresponding to said first RDMA Read request.

13. A method of tracking responses to an RDMA Read operation, the
10 method comprising:
issuing an RDMA Read on a first communication connection;
identifying a range of sequence numbers to be associated with responses to the RDMA Read;
adding an entry to a first linked list corresponding to the first
15 communication connection, said entry configured to identify:
said range of sequence numbers; and
a latest sequence number received in said range of sequence numbers;
receiving a first RDMA Read response;
20 determining whether a first sequence number associated with the first RDMA Read response matches a last sequence number in said range of sequence numbers; and
if said first sequence number does not match said last sequence number, updating said latest sequence number to match said first sequence number.

25

14. The method of claim 13, further comprising:
if said first sequence number matches said last sequence number, retiring

an entry in a retry queue corresponding to the RDMA Read.

15. The method of claim 13, wherein the first communication connection is an InfiniBand queue pair.

5

16. The method of claim 15, wherein:
said issuing is performed by an InfiniBand transmit module; and
said adding, said determining and said updating are performed by an InfiniBand receive module;

10

the method further comprising:

at the InfiniBand transmit module, retrying the RDMA Read if an RDMA Read response associated with said range of sequence numbers is received out of order.

15

17. The method of claim 15, wherein:
said issuing is performed by an InfiniBand transmit module; and
said adding, said determining and said updating are performed by an InfiniBand receive module;

the method further comprising, at the InfiniBand transmit module:

20

maintaining a retry queue for tracking RDMA Reads that have not yet completed; and

retiring an entry in said retry queue corresponding to the RDMA Read if RDMA Read responses corresponding to said range of sequence numbers are received in order.

25

18. The method of claim 17, further comprising, at the InfiniBand transmit module:

retrying the RDMA Read if RDMA Read responses corresponding to one or more of said range of sequence numbers are received out of order.

19. The method of claim 13, wherein said identifying comprises:
5 dividing an amount of data to be received in response to the RDMA Read by a maximum transfer unit in effect for the first communication connection.

20. The method of claim 13, further comprising:
maintaining a single memory structure comprising multiple linked lists
10 corresponding to multiple communication connections, including said first linked list corresponding to the first communication connection.

21. The method of claim 20, further comprising:
for each of the multiple communication connections, including the first
15 communication connection, maintaining pointers to the first entry and the last entry in the corresponding linked list.

22. An apparatus for queuing multiple types of receive traffic in a communication interface, comprising:
20 a queue for queuing multiple types of receive traffic associated with communications to be transmitted from the communication interface;
a head pointer configured to identify a head of said queue;
a tail pointer configured to identify a tail of said queue, wherein said traffic commands are enqueued at said tail;
25 a next entry pointer configured to identify a next entry in said queue to be processed; and
a linked list, wherein each entry in said linked list corresponds to an

RDMA Read request issued by the communication interface, and is configured to identify a range of sequence numbers associated with expected responses to the RDMA Read request.

5 23. The apparatus of claim 22, wherein each entry said linked list is further configured to identify a sequence number of a most recently received response to the RDMA Read request.

 24. The apparatus of claim 22, wherein the linked list is one of
10 multiple linked lists, each said linked list corresponding to a separate InfiniBand queue pair.

 25. The apparatus of claim 22, further comprising:
 a retry queue, wherein a retry entry is added to said retry queue for each
15 RDMA Read request issued by the communication interface;
 wherein a first retry entry in said retry queue corresponding to a first RDMA Read request is retired when said expected responses to the first RDMA Read request are received.

20 26. The apparatus of claim 22, further comprising:
 a memory configured to store pointers to a first entry and a last entry in said linked list.

 27. The apparatus of claim 22, wherein said queue comprises an
25 assembly area for assembling a communication associated with a first type of receive traffic.

28. The apparatus of claim 27, wherein said assembly area comprises a portion of said queue delimited by said head pointer and said next entry pointer.

29. The apparatus of claim 27, wherein said first type of receive traffic
5 is an InfiniBand RDMA Read command comprising a set of RDMA read descriptors configured to identify the communication associated with said first type of receive traffic.

30. The apparatus of claim 29, wherein a second type of receive traffic
10 is an InfiniBand Send command configured to encapsulate the communication associated with said second type of receive traffic command.

31. The apparatus of claim 27, wherein:
said first type of receive traffic comprises a set of descriptors, wherein
15 each said descriptor is configured to describe a portion of the communication associated with said first type of receive traffic; and
the apparatus is configured to issue read requests to retrieve the portions of the communication described by the set of descriptors and assemble said portions in said assembly area.

20

32. The apparatus of claim 22, further comprising:
a transmit module configured to transmit the communications associated with said receive traffic;
wherein each communication associated with receive traffic is forwarded
25 from said queue to said transmit module after the communication is determined to be complete.

33. The apparatus of claim 32, wherein a communication is forwarded from said queue to said transmit module by passing to the transmit module a set of pointers delimiting the communication within said queue.

5 34. The apparatus of claim 22, wherein said queue comprises a linked list of buffers within a memory structure configured to queue receive traffic for multiple communication connections.

35. A communication interface for tracking responses to an InfiniBand
10 RDMA Read request, comprising:

for each of one or more active InfiniBand queue pairs, a corresponding linked list, wherein each entry in said linked list is configured to include:

a range of sequence numbers associated with expected responses to
an RDMA Read request issued on the corresponding queue pair by the
15 communication interface; and

a previous sequence number, wherein said previous sequence number is a sequence number associated with a most recently received response to the RDMA Read request; and

for each of the linked lists, pointers to a first entry and a last entry in said
20 linked list.

36. The communication interface of claim 35, further comprising:
a retry queue configured to queue retry entries corresponding to RDMA
Read requests issued by the communication interface;

25 wherein an retry entry in said retry queue is retired when a final response to a corresponding RDMA Read request is received, said final response being identified by a final sequence number in said range of sequence numbers.

37. The communication interface of claim 35, further comprising:
a transmit module configured to:
issue a first RDMA Read request on a first queue pair; and
5 calculate said range of sequence numbers associated with said
expected responses to the first RDMA Read request; and
a receive module configured to add an entry, corresponding to the first
RDMA Read request, to said corresponding first linked list.
- 10 38. The communication interface of claim 37, wherein said receive
module is further configured to:
determine a sequence number of a response to the first RDMA Read
request; and
determine if said sequence number matches a final sequence number in
15 said range of sequence numbers associated with expected responses to the first
RDMA Read request.
39. The communication interface of claim 38, wherein said receive
module is further configured to:
20 determine if said sequence number is out of order.